London Bike Sharing

Visualization and Exploratory Data Analysis (SVM and Decision Tree)

Thomas Januardy
Information System Major, Faculty of Information & Technology, Multimedia Nusantara University,
Tangerang City, Indonesia
thomas.januardy@student.umn.ac.id

***Abstract*—Data is one of the most important things in the age of all technology as it is today to help the continuity of a process. With a set of data called "big data", we can process an information that helps us in many ways, such as decision making to supporting media. One of the processes that use data in daily activities is the bicycle sharing system using a rental system. To help management and prevent the occurrence of unwanted things in the bicycle rental system, a data system can be implemented with a track-record method, which records activities that occur during the use of one form of object or tool. With that said, the authors wanted to try to calculate how many bike users rent and share bikes in a city based on a given season. Whether there is an increase or decrease if it is in a certain season or not.In addition, from this can also be predicted the use of bicycles in the future based on the seasons in the city. Users use data mining and exploratory methods against the bike sharing data and apply support vector machine and decision tree algorithms to see accuracy and predict bike usage based on the seasons. It is useful to guess patterns and look at the most dominant frequency of bike use among certain seasons, to find out the most frequent and most frequent bike use in a given city.**

Index Terms* — *Bike Sharing, London, Big Data, Data Mining, SVM, Support Vector Machine, Decision Tree, Plot, Visualization, Classification

I. Introduction

A. Big Data Concept

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves [3]. "Big Data" is an aspect that is often needed by large companies today as a very complex and useful accumulation of data for companies. With internet connectivity that is increasingly widespread and fast, we are very easy to get a large collection of informations related to almost every aspect of our lives. The term "big data" refers to data that is so large, fast or complex that it's difficult or impossible to process using traditional methods. The act of accessing and storing large amounts of information for analytics has been around a long time. But the concept of big data gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three V's:

- Volume: Organizations collect data from a variety of sources, including business transactions, smart (IoT) devices, industrial equipment, videos, social media and more. In the past, storing it would have been a problem – but cheaper storage on platforms like data lakes and Hadoop have eased the burden.

- Velocity: With the growth in the Internet of Things, data streams in to businesses at an unprecedented speed and must be handled in a timely manner. RFID tags, sensors and smart meters are driving the need to deal with these torrents of data in near-real time.

- Variety: Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, emails, videos, audios, stock ticker data and financial transactions [3].

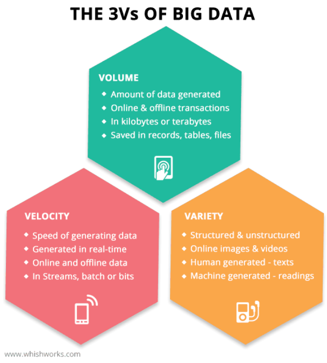


Figure 1. The "3Vs" of Big Data

Big companies like Amazon, Microsoft, Facebook, Google, Twitter, and many others know and understand everything from our travels to our bedtime and our buying patterns. Many believe that, for companies that get it right, big data will be able to unleash new organizational capabilities and value [1].

The creation and collection of data in today's society is not only limited to functions related to finance and consumers. In short, these companies can have a much more complete picture of their customers and operations by combining unstructured and structured data [2].

The “Big Data” era has produced a massive trend, with large organizations striving to leverage big data analytics to create value for these enterprises. Following the rapid expansion of data volume, speed, and variety, substantial developments have been documented in terms of technical and technological developments for data storage, analysis, and visualization. Organizations and enterprises need to change to embrace this innovation, and what business value can be derived from it by data analysis. Businesses must adapt to the ever-changing technological climate data creation and collection in today's society this is not only limited to financial and consumer related functions.

Not only used in companies, data management can also be used for small activities, to find patterns or predict future possibilities. There are many uses for data if it can be managed properly. Companies can search for strategies and make decisions based on available data, a student can research a study on certain people's habits, even predictions of natural disasters that are likely to occur based on time.

B. *Problems*

In this case, there is a vehicle sharing activity in the form of bicycles for rentals in the city of London. Bicycles are one of the most common forms of transportation and are often used by people in the city of London. Therefore, it is not surprising that there are many business and business opportunities in the form of rental with a vehicle (bicycle) sharing system to consumers who need it in the city of London. As we also know that the European continent, where the city of London is located, has more diverse seasons, such as winter, spring, summer, and autumn. Mobility and community activities are also affected and always adapt as the current season progresses, especially the use of bicycles as their daily transportation. From the data collected, the purpose of this study is to predict how much bicycle sharing will occur in the future based on the seasons that are currently being passed in the city of London.

Moreover, as time goes by, vehicles are also increasingly developed and diverse. But a vehicle with a type such as a bicycle, seems to never be rushed by time, because of its simple use and with a mechanism that does not require an engine to be driven. Cycling vehicles can be one solution, where bicycles are vehicles that do not require energy, but humans are the movers themselves. However, vehicles such as bicycles will not always be able to operate if the terrain is not supportive. As we know that bicycles use a rotating wheel system as their route, but the seasons in the city of London, especially the European continent are not as simple as the seasons in Asia and its surroundings. Seasons such as snow and autumn can be one of the obstacles to vehicles such as bicycles when they want to be used. But for some reasons, this does not make it an excuse not to use a bicycle. This is what makes the bicycle categorized as the simplest vehicle and does not require many requirements to be used. For this reason, the research that we want to make as an object is about how the level of bicycle use in rental places is related to the seasons that occur in the city of London.

II. Literature Review

A. *Data Mining and Classification*

Data mining is the process of discovering interesting knowledge from these large amounts of data stored in database, data warehouse or other information repositories. A number of data mining techniques have already been done on educational data mining to improve the performance of students like Regression, Genetic algorithm, Bays classification, k-means clustering, associate rules, prediction etc. Data mining techniques can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students. Classification is one of the most frequently studied problems by data mining and machine learning researchers(Brijesh et al,2011). It consists of predicting the value of a categorical attribute based on the value of other attributes. Classification methods like decision trees, rule mining, Bayesian network etc. can be applied on the educational data for predicting the students behavior, performance in examination etc [4].

B. *Support Vector Machine*

The Support Vector Machine (SVM) was first proposed by Vapnik and has since attracted a high degree of interest in the machine learning research community [7]. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. Sims have been employed in a wide range of real world problems such as text categorization, hand-written digit recognition, tone recognition, image classification and object detection, micro-array gene expression data analysis, data classification. However, for some datasets, the performance of SVM is very sensitive to how the cost parameter and kernel parameters are set. As a result, the user normally needs to conduct extensive cross validation in order to figure out the optimal parameter setting. This process is commonly referred to as model selection. One practical issue with model selection is that this process is very time consuming[8].

SVMs are set of related supervised learning methods used for classification and regression [7]. They belong to a family of generalized linear classification. A special property of SVM is , SVM simultaneously minimize the empirical classification error and maximize the geometric margin.

SVM has pros and cons that associated with SVM itself. The pros are:

- It works really well with a clear margin of separation
- It is effective in high dimensional spaces.
- It is effective in cases where the number of dimensions is greater than the number of samples.
- It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

And the cons are:

- It doesn't perform well when we have large data set because the required training time is higher
- It also doesn't perform very well, when the data set has more noise i.e. target classes are overlapping
- SVM doesn't directly provide probability estimates, these are calculated using an expensive five-fold cross-validation. It is included in the related SVC method of Python scikit-learn library [9].

C. *Decision Tree*

Decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles and the leaf nodes are denoted by ovals. It the most commonly used algorithm because of its ease of implementation and easier to understand compared to other classification algorithms (Surjeet kumar et al,2012).

Decision tree can be constructed relatively fast compared to other methods of classification. Trees can be easily converted into SQL statements that can be used to access databases efficiently. Decision tree classifiers obtain similar and sometimes better accuracy when compared with other classification methods. Decision tree algorithm can be implemented in a serial or parallel fashion based on the volume of data, memory space available on the computer resource and scalability of the algorithm [4].

First introduced in 1960's, decision trees are one of the most effective methods for data mining; they have been widely used in several disciplines [5] because they are easy to be used, free of ambiguity, and robust even in the presence of missing values [6].

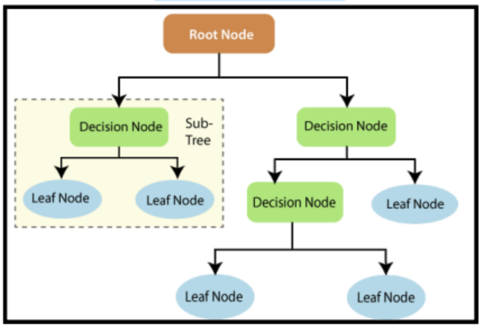


Figure 2. The Model Structure of Decision Tree

Common uses of decision tree models include:

- Variable Selection
- Assesing the relative importance of variables
- Handling of missing values
- Prediction
- Data manipulation

Decision tree method is a powerful statistical tool for classification, prediction, interpretation and data processing, and it has many potential applications in medical research. Using decision tree models to describe research results has the following advantages:

- Simplifies complex relationships between input variables and target variables by dividing original input variables into significant subgroups.
- Easy to understand and interpret.
- Non-parametric approach without distributional assumptions.
- Easy to handle missing values without needing to resort to imputation.
- Easy to handle heavy skewed data without needing to resort to data transformation.
- Robust to outliers.

As with all analysis methods, users must pay attention to the limitations of the decision tree method. The main disadvantage is that it may overfit and underfit, especially when using small data sets. This problem will limit the universality and robustness of the resulting model. Another potential problem is that the strong correlation between different potential input variables may lead to selected variables that can improve model statistics, but have no causal relationship with the results of interest. Therefore, care must be taken when interpreting decision tree models and using the results of these models to formulate causal assumptions.

III. Methodology

A. *Object of Research*

The object of research that we use as material for the first research is the large number of people who rent bicycles specifically in the city of London based on the season and the temperature in the seasons in the city of London. These seasons include "Spring", "Summer", "Fall", and "Winter". All data has been recorded regarding the time, date, until season that is being experienced when renting a bicycle in the city of London.

The dataset has met the minimum criteria, specified for research purposes in the "Data Analysis" course with 10 variable columns and 17,415 data observations, including the headers of each variable. The contents of the observation include data on bicycle rental activities that were recorded and occurred at that time. While the variable column contains timestamp (char), cnt (int), t1 (temperature/num), t2 (temperature feels/num), hum (humidity/num), wind_speed (num), weather_code (num), is_holiday (num) , is_holiday (num), and season (num). Some variables such as weather_code, is_holiday, is_weekend, and season need to be factored into categorical because they are not numeric data, but categorical. Although the target variable is based on the season, the season variable is a priority that needs to be factored into categorical data. The timestamp variable also needs to be changed to the datetime data type so that the date data can be identified. Apart from that, they are numeric datas that contain numbers.

B. Method of Collecting Data

Our collection method that we did is by retrieving a dataset from one of the well-known dataset search websites called “Kaggle”. This is because the website has much experienced online dataset, especially in science community and machine learning practitioners. Kaggle also allows us to discover and publish power sets, explore and build models in web-based data science environments, work with data scientists and other machine learning engineers, and enter competitions to solve data science challenges. The reason we also take the dataset in the Kaggle application is because the completeness of the content of the data from the website is very complete and freely accessed. Any people can access the data and can process it. Kaggle has many helpful features such as notebooks as a means for users to display their work data.

C. Research Methods

Research method is a step taken in order to collect information or data and evaluate the data that has been obtained. Research methods provide an overview of the research design that includes research procedures and steps to be done, research time, data sources, and by what steps the data is obtained and the next stage is processed and analyzed to get a conclusion. Because the data that want to be looked for is numerical, the method that can be offered is the management of quantitative data, which is used as data to predict the use of bicycle sharing based on the seasons that occur in the city of London.

The program that will be used to manage datasets is R. R is a programming language. programming language as well as a computing program used to support statistical and graph analysis activities. R accessed using RStudio. RStudio is the Integrated Development Environment (IDE) for R.

Before getting into the algorithm, we need to explore and explore the selected dataset. This needs to be done to find out the outside and in the data, whether the selected dataset will be the optimal data to apply data analysis or not, whether the dataset matches the algorithm, and so on. Therefore, it is necessary to make observations with data mining in the dataset. Data mining is an extension of traditional data analysis and statistical approaches in that it incorporates analytical techniques drawn from a range of disciplines including, but not limited to,

- numerical analysis,
- pattern matching and areas of artificial intelligence such as machine learning,
- neural networks and genetic algorithms.

While many data mining tasks follow a traditional, hypothesis-driven data analysis approach, it is commonplace to employ an opportunistic, data driven approach that encourages the pattern detection algorithms to find useful trends, patterns, and relationships [11].

From the overall data, we will look at how much bicycle rental activities occur in the city of London based on all the seasons in the future. In looking for such data, we will use 2 algorithms to find the accuracy that will be used to predict the purpose of this research. Both algorithms are the right algorithms for classification data, because first, decision tree is a simple and easy algorithm to understand for beginners, and support vector machine, is an efficient algorithm that involves vectors for many applications in science and engineering, especially for classification problems [10], and predict a pattern and is very suitable for this type of classification data.

IV. Results and Discussion

A. Data Validation

Before exploring and processing the data, it is necessary to check the validity of the datas. What is meant to be validity is a condition of the overall observations contained in the dataset. Whether there is data that is damaged or lost, or can also check the existence of duplicated data. This needs to be straightened out first before starting the first data analysis step because checking is an essential step in processing data. One is by way of whether all variable columns have actual data with the other or still exist incomplete, using the "missmap" function in R.T he result is that no data is lost at all and the data is one hundred percent complete. Thus, it can be concluded that the data has been valid for use in research because it has passed the stage of examination of the truth and completeness of the data.

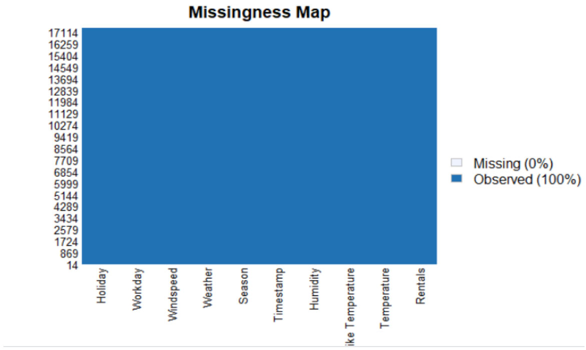
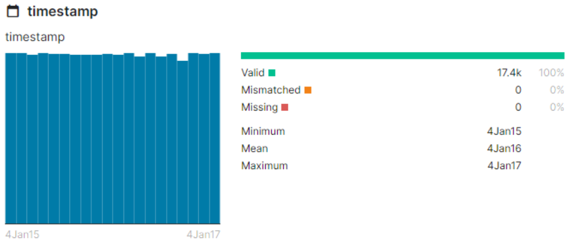


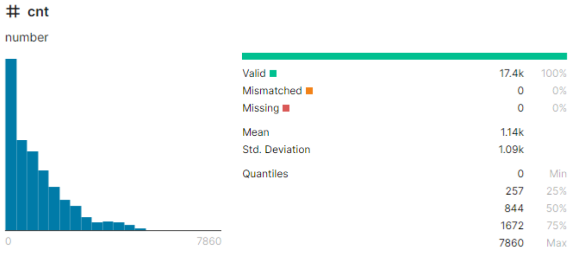
Figure 3. Missingness Map

As for the examination of data validity through Kaggle itself to check the completeness of observation data in each column of variables that the dataset provides. Of the 10 variables, no data was found. It can be concluded that all variables are safe to use and ready to be processed to the next stage. Here is the breakdown of the variable columns from Kaggle:



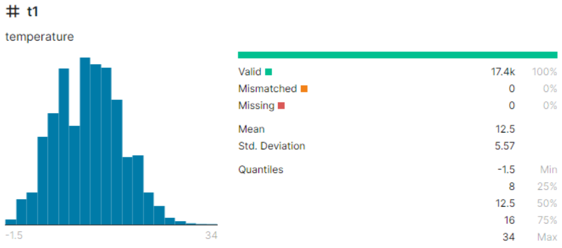
```
timestamp
timestamp

Valid          17.4k  100%
Mismatched         0    0%
Missing            0    0%

Minimum        4Jan15
Mean           4Jan16
Maximum        4Jan17
```

Figure 4. “timestamp” variable



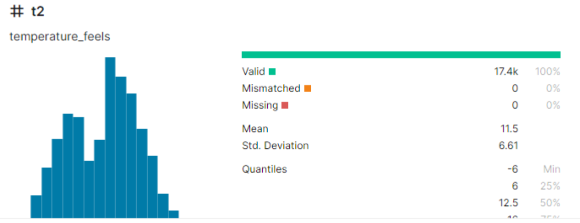
```
# cnt
number

Valid           17.4k  100%
Mismatched          0    0%
Missing             0    0%

Mean            1.14k
Std. Deviation  1.09k

Quantiles           0   Min
                  257   25%
                  844   50%
                 1672   75%
                 7860   Max
```

Figure 5. “cnt” variable



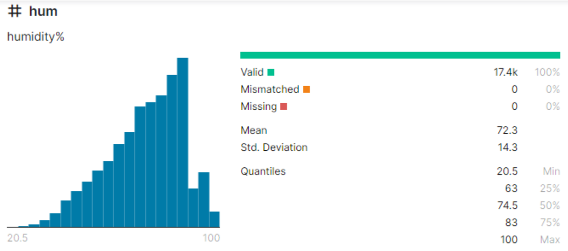
```
# t1
temperature

Valid           17.4k  100%
Mismatched          0    0%
Missing             0    0%

Mean             12.5
Std. Deviation   5.57

Quantiles        -1.5   Min
                    8   25%
                 12.5   50%
                   16   75%
                   34   Max
```

Figure 6. “t1” variable



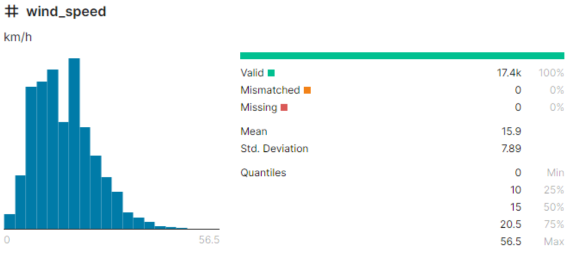
```
# t2
temperature_feels

Valid           17.4k  100%
Mismatched          0    0%
Missing             0    0%

Mean             11.5
Std. Deviation   6.61

Quantiles          -6   Min
                    6   25%
                 12.5   50%
```

Figure 7. “t2” variable



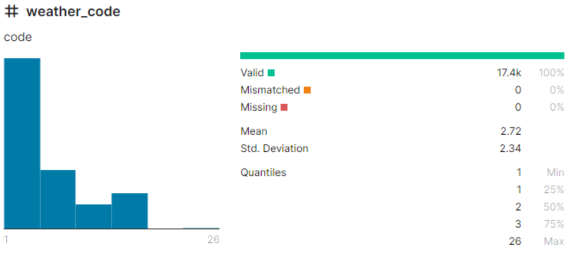
```
# hum
humidity%

Valid           17.4k  100%
Mismatched          0    0%
Missing             0    0%

Mean             72.3
Std. Deviation   14.3

Quantiles        20.5   Min
                   63   25%
                 74.5   50%
                   83   75%
                  100   Max
```

Figure 8. “hum” variable



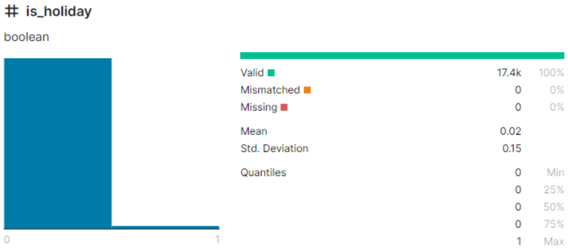
```
# wind_speed
km/h

Valid           17.4k  100%
Mismatched          0    0%
Missing             0    0%

Mean             15.9
Std. Deviation   7.89

Quantiles           0   Min
                   10   25%
                   15   50%
                 20.5   75%
                 56.5   Max
```

Figure 9. “wind_speed” variable



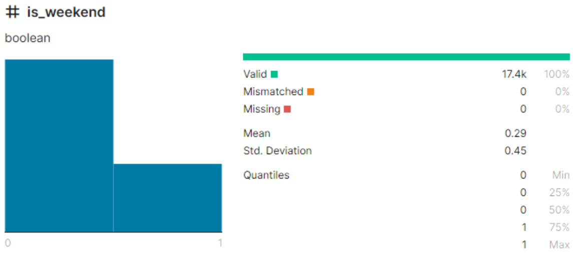
```
# weather_code
code

Valid           17.4k  100%
Mismatched          0    0%
Missing             0    0%

Mean             2.72
Std. Deviation   2.34

Quantiles           1   Min
                    1   25%
                    2   50%
                    3   75%
                   26   Max
```

Figure 10. “weather_code” variable



```
# is_holiday
boolean

Valid           17.4k  100%
Mismatched          0    0%
Missing             0    0%

Mean             0.02
Std. Deviation   0.15

Quantiles           0   Min
                    0   25%
                    0   50%
                    0   75%
                    1   Max
```

Figure 11. “is_holiday” variable

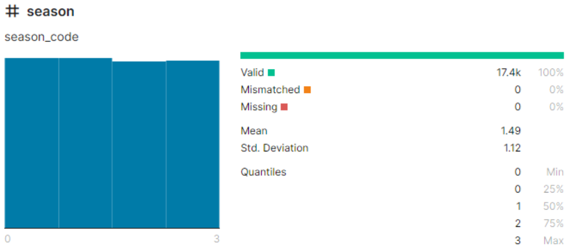


is_weekend

boolean

Valid	17.4k	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.29	
Std. Deviation	0.45	
Quantiles	0	Min
	0	25%
	0	50%
	1	75%
	1	Max

Figure 12. “is_weekend” variable



season

season_code

Valid	17.4k	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.49	
Std. Deviation	1.12	
Quantiles	0	Min
	0	25%
	1	50%
	2	75%
	3	Max

Figure 13. “season” variable

B. *Data Visualization*

1) *Visualization: Rentals*

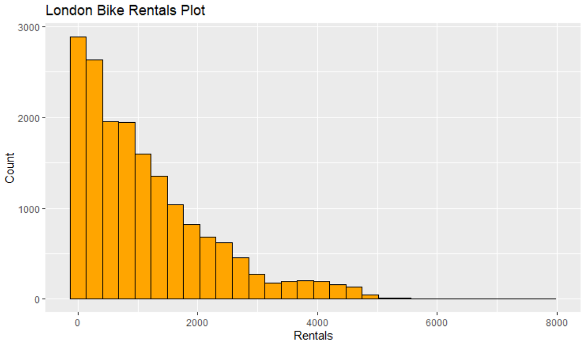


Figure 14. Visualization: Rentals

2) *Visualization: Rentals by Season Barplot*

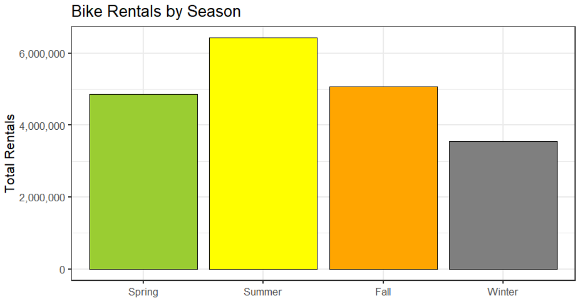


Figure 15. Visualization: Rentals by Season Barplot

3) *Visualization: Rentals by Season*

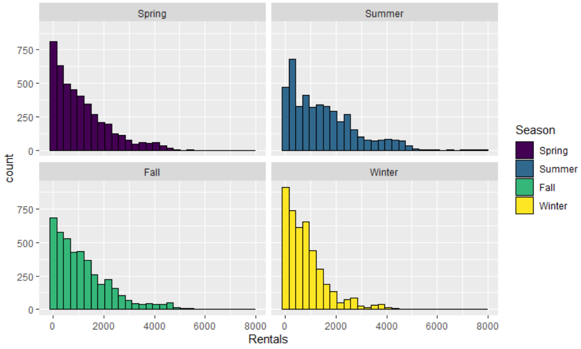


Figure 16. Visualization: Rentals by Season

4) *Visualization: Rentals by Season Boxplot*

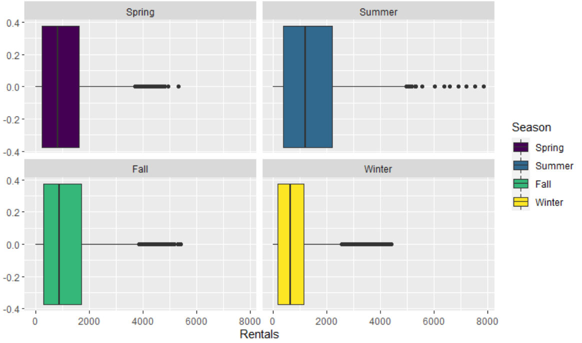


Figure 17. Visualization: Rentals by Season Boxplot

5) *Visualization: Rentals by Season (with Temperatures)*

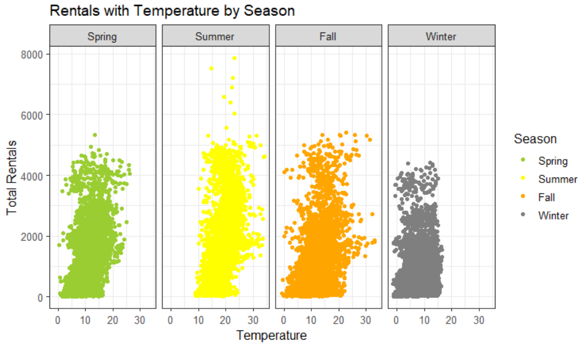
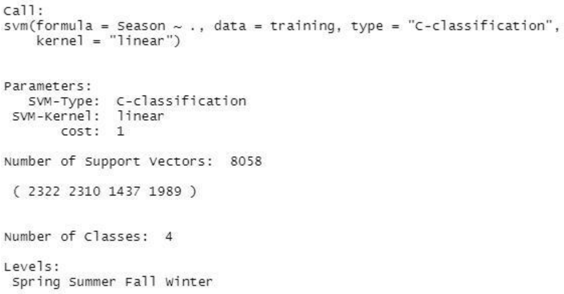


Figure 18. Visualization: Rentals by Season (with Temperatures)

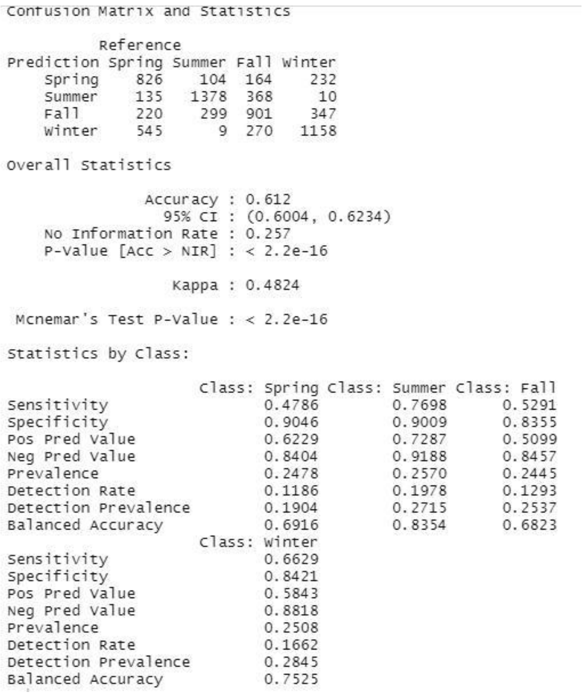
C. *Algorithms*

1) *Support Vector Machine*



```
Call:
svm(formula = Season ~ ., data = training, type = "C-classification", 
    kernel = "linear")


Parameters:
   SVM-Type:  C-classification 
 SVM-Kernel:  linear 
       cost:  1 

Number of Support Vectors:  8058

 ( 2322 2310 1437 1989 )


Number of Classes:  4 

Levels: 
 Spring Summer Fall Winter
```

Figure 19. Support Vector Machine on “London” Data



```
Confusion Matrix and Statistics

          Reference
Prediction Spring Summer Fall Winter
    Spring    826    104  164    232
    Summer    135   1378  368     10
    Fall      220    299  901    347
    Winter    545      9  270   1158

Overall Statistics

               Accuracy : 0.612
                 95% CI : (0.6004, 0.6234)
    No Information Rate : 0.257
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4824

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: Spring Class: Summer Class: Fall
Sensitivity                 0.4786        0.7698      0.5291
Specificity                 0.9046        0.9009      0.8355
Pos Pred Value              0.6229        0.7287      0.5099
Neg Pred Value              0.8404        0.9188      0.8457
Prevalence                  0.2478        0.2570      0.2445
Detection Rate              0.1186        0.1978      0.1293
Detection Prevalence        0.1904        0.2715      0.2537
Balanced Accuracy           0.6916        0.8354      0.6823
                     Class: Winter
Sensitivity                 0.6629
Specificity                 0.8421
Pos Pred Value              0.5843
Neg Pred Value              0.8818
Prevalence                  0.2508
Detection Rate              0.1662
Detection Prevalence        0.2845
Balanced Accuracy           0.7525
```

Figure 20. SVM's Confusion Matrix on "Testing" Newdata

SVM calculations using the dataset "london" with the target variable "Season" gets vectors as much as 8,058 with c-classification type and linear kernel. In splitting data, the comparison is done by 60:40 between training and testing data, so that the nrow that tests produce only 6,966 while training is 10,448. SVM fittings use newdata from training to match the entire data.

In the prediction of data, from both newdata existing, is much better testing than training. This is because newdata testing has greater accuracy than training although not too large, but quite distinguishing. The confusion matrix value produces an accuracy in newdata testing of 0.612 or 61% to the variable "Season". Seen from the sensitivity to the classes "Season" can be distinguished the most dominant data distribution from others and most impacting this bike sharing activity, is the "Summer" class of 0.76 or 76% of data.

2) *Decision Tree*

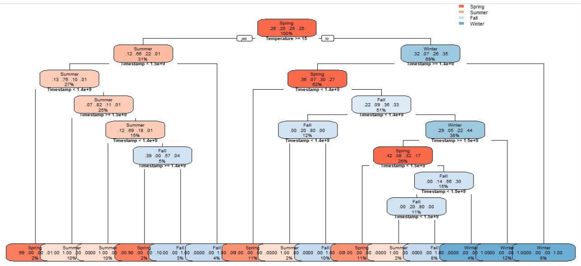
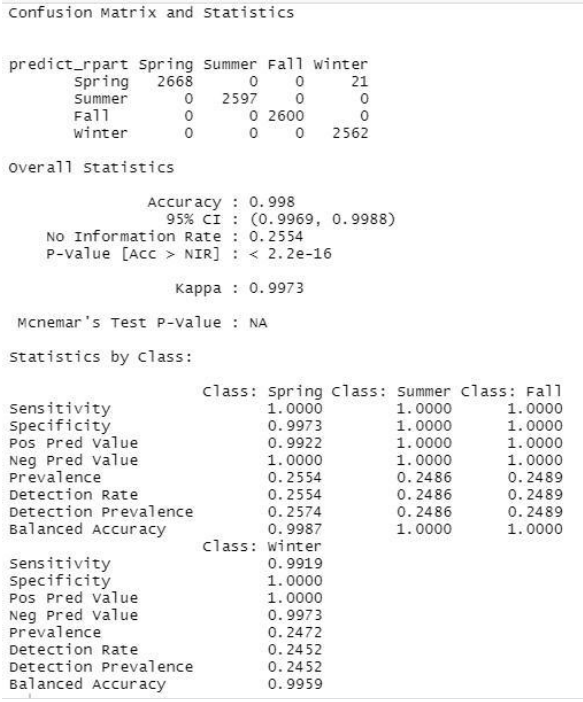


Figure 21. Rpart Plot on "Training2" Newdata

The Decision Tree algorithm that used for this dataset uses from the rpart package, because the rpart package is the appropriate and provides maximum results than other packages, such as package "party". Decision tree rpart is the entire data of the occurrence of thickening that occurs in a particular "Season" and is based on "Timestamp" and "Temperature, where the time and temperature that occurs during the use of bicycles in the city of London, refers to the decision node.There are 14 selection processes (if) in this decision tree rpart plot, referring to the variable "Season" which contains 4 classes (Spring, Summer, Fall, Winter).



```
Confusion Matrix and Statistics


predict_rpart Spring Summer Fall Winter
       Spring   2668      0    0     21
       Summer      0   2597    0      0
       Fall        0      0 2600      0
       Winter      0      0    0   2562

Overall Statistics

               Accuracy : 0.998
                 95% CI : (0.9969, 0.9988)
    No Information Rate : 0.2554
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9973

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Spring Class: Summer Class: Fall
Sensitivity                 1.0000        1.0000      1.0000
Specificity                 0.9973        1.0000      1.0000
Pos Pred Value              0.9922        1.0000      1.0000
Neg Pred Value              1.0000        1.0000      1.0000
Prevalence                  0.2554        0.2486      0.2489
Detection Rate              0.2554        0.2486      0.2489
Detection Prevalence        0.2574        0.2486      0.2489
Balanced Accuracy           0.9987        1.0000      1.0000
                     Class: Winter
Sensitivity                 0.9919
Specificity                 1.0000
Pos Pred Value              1.0000
Neg Pred Value              0.9973
Prevalence                  0.2472
Detection Rate              0.2452
Detection Prevalence        0.2452
Balanced Accuracy           0.9959
```

Figure 22. Rpart's Confusion Matrix on "Training2" Newdata

The accuracy of this rpart decision tree model is 0.998 or 99%, which is close to the value of 1. Using newdata "Training" with a ratio of 60:40 with newdata "Testing", gives the result that newdata "Training" is the best sample when splitting data because it has a greater accuracy rate than newdata "Testing" itself. There are 14 selection processes in this plot, and have conditions that support the variables of the "Season" itself, including "Timestamp" (dating) and "Temperature" (the temperature that is happening).The fourteen selection processes that occur on the plot, can be seen further in figure 20.

V. Conclusion

Based on the results and information on each algorithm, it can be concluded that the decision tree algorithm is the best algorithm for the dataset this time because of its accuracy level that provides near-perfect values compared to the SVM algorithm in this "London Bike Sharing" dataset. From the results of SVM algorithm calculations that have an accuracy output of 61%, it is still far inferior to the accuracy of the decision tree algorithm with an accuracy output of 99%. This provides an explanation that the decision tree algorithm (rpart) is appropriate and can be used in this dataset.

Confusion Matrix	Support Vector Machine	Decision Tree
Accuracy	0.612	0.998
95% CI	(0.6004, 0.6234)	(0.9969, 0.9988)
Kappa	0.4824	0.9973

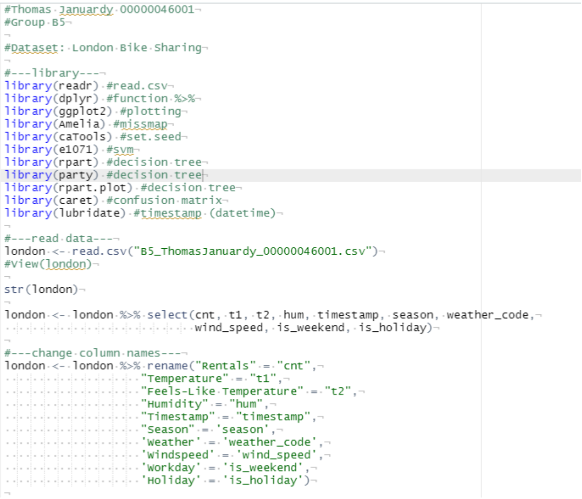
Table 1. SVM and Decision Tree's Comparisons

Based on the comparison table of the two algorithms above, there is a significant difference between svm algorithm and decision tree. The decision tree algorithm delivers results that are almost all close to the perfect value, which is 1, ranging from accuracy of 0.998, 95% CI of (0.9969, 0.9988), and kappa of 0.9973, from confusion matrix assessment. In contrast to SVM which gives an average value in the numbers 0.5 and 0.6, which only offers results from 3/4 of the decision tree algorithm, with an accuracy of 0.612.95% CI of (0.6004, 0.6234), and kappa of 0.4824.

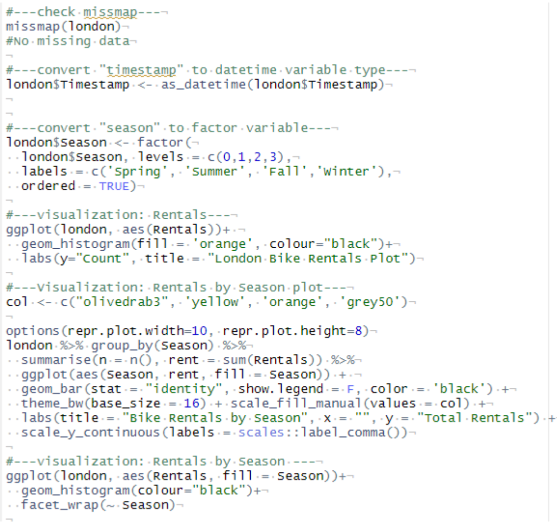
With this research, it can be concluded that the most crowded bicycle use is in the "Summer" season, with the criteria of the season that is very supportive for transportation use such as this bike. But that does not rule out the possibility that in other seasons there will be a significant decrease, besides in other seasons there is also a not too distant use between these seasons, approximately 4 million bicycle uses from the entire season, with the season "Winter" as the lowest observation. As we also know, that season like "Winter" will make it difficult for motorists, especially on land vehicles that use wheels. Bicycles will also be difficult to use if the path through which it passes does not fit the criteria for using wheels on certain roads, such as due to snow or closed and dirty roads. From the predictions displayed also want to inform that the use of bicycles can continue to increase from year to year over time. This happens because bicycles are a simple and healthy transportation.In addition to being used for sports, bicycle transportation can also support environmental sustainability by not using transportation that produces emissions at all. But keep in mind also that this research is just a task workmanship and still has many mistakes and shortcomings. All data management and management is not necessarily true with the facts. It is expected that in the future it can be improved and developed again with more diverse data as a task workmanship and portfolio that is useful for others.

Attachments

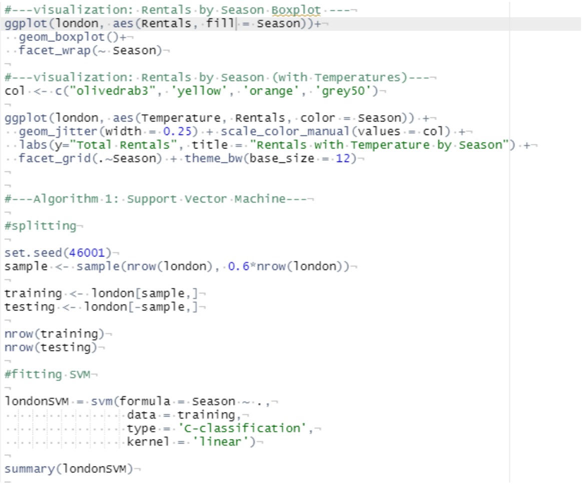
Here is the R code that has been created in the workmanship and management of data in the dataset "London Bike Sharing" by Thomas Januardy (00000046001):



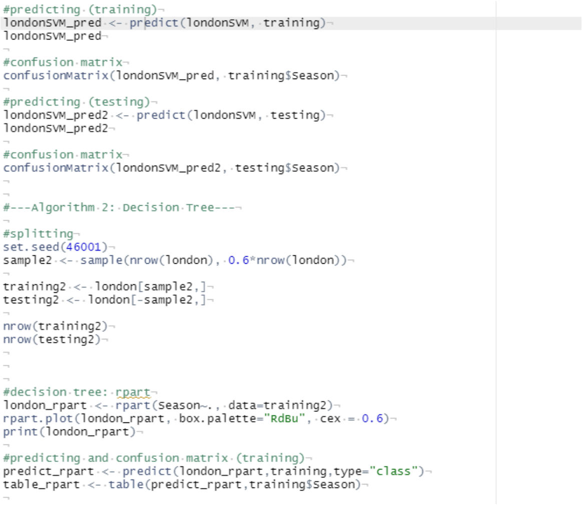
```r
#Thomas Januardy 00000046001
#Group B5

#Dataset: London Bike Sharing

#---library---
library(readr) #read.csv
library(dplyr) #function %>%
library(ggplot2) #plotting
library(Amelia) #missmap
library(caTools) #set.seed
library(e1071) #svm
library(rpart) #decision tree
library(party) #decision tree
library(rpart.plot) #decision tree
library(caret) #confusion matrix
library(lubridate) #timestamp (datetime)

#---read data---
london <- read.csv("B5_ThomasJanuardy_00000046001.csv")
#View(london)

str(london)

london <- london %>% select(cnt, t1, t2, hum, timestamp, season, weather_code,
                            wind_speed, is_weekend, is_holiday)

#---change column names---
london <- london %>% rename("Rentals" = "cnt",
                      "Temperature" = "t1",
                      "Feels-Like Temperature" = "t2",
                      "Humidity" = "hum",
                      "Timestamp" = "timestamp",
                      "Season" = 'season',
                      'Weather' = 'weather_code',
                      'Windspeed' = 'wind_speed',
                      'Workday' = 'is_weekend',
                      'Holiday' = 'is_holiday')

```



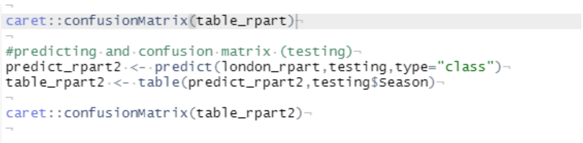
```r
#---check missmap---
missmap(london)
#No missing data

#---convert "timestamp" to datetime variable type---
london$Timestamp <- as_datetime(london$Timestamp)


#---convert "season" to factor variable---
london$Season <- factor(
  london$Season, levels = c(0,1,2,3),
  labels = c('Spring', 'Summer', 'Fall','Winter'),
  ordered = TRUE)

#---visualization: Rentals---
ggplot(london, aes(Rentals))+ 
  geom_histogram(fill = 'orange', colour="black")+
  labs(y="Count", title = "London Bike Rentals Plot")

#---Visualization: Rentals by Season plot---
col <- c("olivedrab3", 'yellow', 'orange', 'grey50')

options(repr.plot.width=10, repr.plot.height=8)
london %>% group_by(Season) %>%
  summarise(n = n(), rent = sum(Rentals)) %>%
  ggplot(aes(Season, rent, fill = Season)) + 
  geom_bar(stat = "identity", show.legend = F, color = 'black') +
  theme_bw(base_size = 16) + scale_fill_manual(values = col) +
  labs(title = "Bike Rentals by Season", x = "", y = "Total Rentals") +
  scale_y_continuous(labels = scales::label_comma())

#---visualization: Rentals by Season ---
ggplot(london, aes(Rentals, fill = Season))+
  geom_histogram(colour="black")+
  facet_wrap(~ Season)

```



```r
#---visualization: Rentals by Season Boxplot ---
ggplot(london, aes(Rentals, fill = Season))+
  geom_boxplot()+
  facet_wrap(~ Season)

#---visualization: Rentals by Season (with Temperatures)---
col <- c("olivedrab3", 'yellow', 'orange', 'grey50')

ggplot(london, aes(Temperature, Rentals, color = Season)) +
  geom_jitter(width = 0.25) + scale_color_manual(values = col) +
  labs(y="Total Rentals", title = "Rentals with Temperature by Season") +
  facet_grid(.~Season) + theme_bw(base_size = 12)


#---Algorithm 1: Support Vector Machine---

#splitting

set.seed(46001)
sample <- sample(nrow(london), 0.6*nrow(london))

training <- london[sample,]
testing <- london[-sample,]

nrow(training)
nrow(testing)

#fitting SVM

londonSVM = svm(formula = Season ~ .,
                data = training,
                type = 'C-classification',
                kernel = 'linear')

summary(londonSVM)

```



```
#predicting (training)
londonSVM_pred <- predict(londonSVM, training)
londonSVM_pred

#confusion matrix
confusionMatrix(londonSVM_pred, training$Season)

#predicting (testing)
londonSVM_pred2 <- predict(londonSVM, testing)
londonSVM_pred2

#confusion matrix
confusionMatrix(londonSVM_pred2, testing$Season)


#---Algorithm 2: Decision Tree---

#splitting
set.seed(46001)
sample2 <- sample(nrow(london), 0.6*nrow(london))

training2 <- london[sample2,]
testing2 <- london[-sample2,]

nrow(training2)
nrow(testing2)



#decision tree: rpart
london_rpart <- rpart(Season~., data=training2)
rpart.plot(london_rpart, box.palette="RdBu", cex = 0.6)
print(london_rpart)

#predicting and confusion matrix (training)
predict_rpart <- predict(london_rpart,training,type="class")
table_rpart <- table(predict_rpart,training$Season)

```



```

caret::confusionMatrix(table_rpart)

#predicting and confusion matrix (testing)
predict_rpart2 <- predict(london_rpart,testing,type="class")
table_rpart2 <- table(predict_rpart2,testing$Season)

caret::confusionMatrix(table_rpart2)

```

Thank-you Note

This paper will never be completed without the help and guidance of Mrs. Tan Thing Heng, Bsc, Mstats as lecturer of the course of "Data Analysis", Information Systems major, Multimedia Nusantara University. The author would like to thank Mrs. Tan Thing Heng for the availability of time and consultation sessions in the work of this task and paper by providing solutions and guidance so it makes and paper can be completed properly.

References

[1] Thomas H. Davenport, Paul Barth, and Randy Bean, “How ‘Big Data’ is Different”, MIT Sloan Management Review, FALL 2012 Vol. 54, No.1, p. 22, 2012.

[2] Thomas H. Davenport, Jill Dyché, “Big Data in Big Companies”, International Institute for Analytics, 2013.

[3] Thomas H. Davenport. 2013 http://www.sas.com/en_th/insights/bigdata/what-is-big-data.html, Accessed 8 October 2021, 23:00.

[4] Anuja Priyam, Abhijeet, Rahul Gupta, Anju Rathee, and Saurabh Srivastava, “Comparative Analysis of Decision Tree Classification Algorithms”, International Journal of Current Engineering and Technology, Vol. 3 No. 2, June 2013.

[5] Hastie TJ, Tibshirani RJ, and Friedman JH, “The Elements of Statistical Learning: Data Mining Inference and Prediction”, Second Edition, Springer; 2009, ISBN 978-0-387-84857-0.

[6] Yan-yan SONG, Ying LU, “Decision tree methods: applications for classification and prediction”, Shanghai Archives of Psychiatry, April 2015.

[7] V. Vapnik, “The Nature of Statistical Learning Theory”, NY: Springer-Verlag, 1995.

[8] Durgesh K. Srivastava, Lekha Bhambhu, “DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE”, Journal of Theoretical and Applied Information Technology, 2005-2009.

[9] Sunil Ray, “Understanding Support Vector Machine(SVM) algorithm from examples (along with code)”, September 2017, Analytics Vidhya, https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/, Accessed 8 October 2021, 23:43.

[10] Qiang Wu, Ding-Xuan Zhou, “Analysis of Support Vector Machine Classification”, Journal of Computational Analysis and Applications, April 2006.

[11] Joyce Jackson, “Data Mining; A Conceptual Overview”, Communications of the Association for Information Systems, Volume 8, Article 19, March 2002.